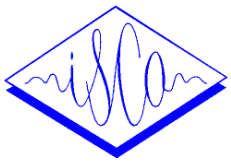

2nd European Conference on
Speech Communication and Technology
EUROSPEECH '91
Genova, Italy, September 24-26, 1991

A MAN-MACHINE DIALOGUE SYSTEM FOR SPEECH ACCESS TO E-MAIL INFORMATION USING THE TELEPHONE: IMPLEMENTATION AND FIRST RESULTS*

P. Baggia A. Ciaramella D. Clementino L. Fissore E. Gerbino E. Giachin
G. Micca L. Nebbia R. Pacifici G. Pirani C. Rullent

CSELT - Centro Studi e Laboratori Telecomunicazioni
Via G. Reiss Romoli 274 - 10148 Torino, Italy

Abstract

In the framework of the SUNDIAL Esprit System we are developing a man-machine dialogue system for interactive speech access to a remote data base. We describe the major system blocks, i.e. the acoustical front-end, the linguistic processor, the dialogue management and the message generation components, as well as their interplay. Finally we give an overview of the first speed and accuracy performance of the real time demonstrator.

1 Introduction

On the basis of the achievements obtained in the framework of the SUNDIAL Esprit project [1] we are developing a man-machine dialogue system for speech access to E-mail information using the telephone.

This is one of the possible applications of interactive telephone speech access to a remote data base: other possible applications are flight or train information access, which are being developed in the SUNDIAL project by the partners in different languages. These applications stress the state of the art technology since they require:

- large vocabulary continuous speech recognition, speaker independent and robust to channel and telephone hand-set variability,
- a natural language understanding capability,
- an intelligent dialogue manager, with mixed user and system initiative,
- a good quality text synthesis through telephone lines.

In the following we will describe the system architecture, the present implementation and the performance results achieved in our E-mail real-time demonstrator.

*This work has been partially supported by CEC Esprit II project 2218 – SUNDIAL:Speech UNderstanding and DIALogue.

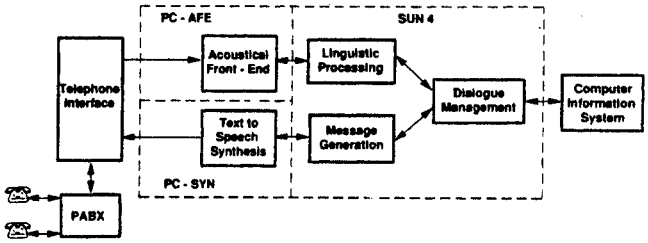

Figure 1: System Block Diagram

2 System architecture

2.1 General overview

The system is composed of 4 major blocks (Fig. 1):

- the acoustical front-end (AFE),
- the linguistic processor (LP),
- the dialogue manager(DM),
- the message generator and synthesis (MG);

for the Italian demonstrator the DM and the MG are devoloped outside of the SUNDIAL project. A telephone interface interconnects the acoustical front-end and the speech synthesis to the telephone network, whilst the dialogue manager accesses the remote data base.

The system has an application vocabulary of 786 lexical entries, of which 117 are function sentences composed of two or more single words; the AFE was trained by means of a 8800 continuous sentence database collected through a PABX line from 110 speakers spanning a wide range of ages and dialectal inflections.

The system is presently housed in 4 cabinets (Fig.1):

- a SUN4/260, containing as separate tasks the LP, the DM with the message generator and the overall system control,

- a PC equipped with DSP boards, implementing the AFE [2],
- a PC equipped with an LPC custom board, implementing the text synthesis,
- the telephone interface, recognizing the telephone calls and disconnections and transforming them in computer manageable form (RS-232 ASCII codes).

The PC implemented AFE is connected to the SUN4 through a pair of communication boards, with a transfer speed of 6 Mbytes/s: this link requires in fact fast communication capabilities.

The PC implemented synthesizer is connected to the SUN4 through an RS-232 serial link: in this case we have to mainly transfer output messages in ASCII form.

2.2 System use

The user and the system have to speak alternatively; at the moment only one case of speech overlap is allowed, letting the user to stop the synthesis of a message with the keyword "FINE" (i.e. "stop") pronounced as an isolated word.

In fact, according to the dialogue evolution, the system can switch from isolated word (IWR) to continuous speech recognition (CSR) and viceversa; whilst CSR is the most frequently used mode, IWR is used in most critical situations, as:

- initial acquisition of user name and password,
- negative or positive confirmation of critical commands, as for example message canceling,
- stop of the synthesis of messages,
- recovery of difficult dialogue sessions, by switching to a menu driven dialogue strategy.

2.3 The acoustical front-end

The acoustical front-end can be distinguished into two sections

- the feature extraction stage,
- the pattern matching stage.

The feature extraction stage performs the following computations:

- end-point detection,
- parameters evaluation,
- vector quantization.

The feature extraction stage computes energy, delta energy, 12 DCTs and 12 delta DCTs [7] of the utterance at a 10 ms frame rate, then performs vector quantization, obtaining three codevectors: the first one for the DCTs, the second one for the delta DCTs, the third one for energy and delta energy; the first two codevectors are of 8 bits, whilst the last of 5 bits.

These codevectors are used as input to the pattern matching stage, which can work in IWR or in CSR mode: in both cases the lexicon words are described as a sequence of acoustic phonetic units (APU) [4], collapsed into an APU labelled words tree; each of the different APUs is represented as a 3-state discrete density hidden Markov model (DDHMM) [3]. A subset of the nodes of the tree correspond to the end of specific words.

The stream of utterance codevectors is used to find the best paths in the selected vocabulary tree: in CSR the recognition iterates through the word tree: when a terminal node is reached, the root node is reentered; in the IWR the word tree can be gone through only once.

In both cases the Forward or the Viterbi algorithms [7] can be used for identifying the most likely paths in the word tree: the Viterbi decoder has been selected for the real-time demonstrator, since it is simpler to implement. Whilst the IW recogniser simply extracts the ordered list of most likely words, the CS recogniser extracts a lattice of most likely words, with the corresponding time limits and scores. Words appear in this lattice if in a given time interval the likelihood of the path from the first to the last state of the word remains above a given threshold [5]. As a final post-processing the AFE calculates the best sequence of words in the lattice: this additional information increases the understanding speed and accuracy, as we will demonstrate in the following. AFE computations are not terminated here in the CSR case, since a feedback verification phase can be required, as detailed in the next paragraph.

2.4 The Linguistic Module

The LP, activated by the presence of the lattice of word hypotheses, parses it using a new score-based, island-driven control strategy [8]. Linguistic knowledge is represented, at the user level, using a Dependency Grammar for sintax and Caseframes for semantics. Dependency rules and caseframes are compiled into structures called Knowledge Sources, which retain the basic structure of the dependency rules and add semantic constraints to their constituents [9]. These structures are fused together in different ways to increase the global parsing efficiency [10]. A characteristic of our LP is that it is able to process a lattice without requiring in it the presence of short function words, which are often unreliably recognized or even missing in the lattice itself [11]. The parser can continue the analysis when a solution is found; in this case we can activate the feedback verification procedure: the set of sentences extracted by the LP, which are hence both syntactically and semantically correct, are sent again to the AFE in order to receive a final acoustical scoring: of course the AFE has retained in its memory all the codevector sequence of the utterance to be verified [2]. This final verification is not a mere duplication of the first AFE lattice extraction, but recognises the utterance in a predefined set of word sequences, converting a CSR to a IWR-like algorithm. This final AFE verification hence improves the recognition accuracy, since not only scores of words are combined in a more natural way than in the lattice, but short function words hypotesized by the LP can be really verified from the acoustical point of view. After this

AFE verification, the LP completes the parsing activities, generating the internal semantic structure.

2.5 The Dialogue Manager and Message Generator

The DM interpretes the meaning of the LP solution in the context of the ongoing dialogue. This context is represented through a dialogue history for the utterances and a context hierarchy for restrictions and focus shifting. For each utterance the DM must solve elliptical and anaphoric references [12], find the resolution context, update the dialogue history and context hierarchy, access the database and finally interact with the MG module. By using a pattern based technique, the MG can generate different kinds of sentences and texts like data description, messages, confirmation for critical operations, recovery sentences, etc..The generated answers contain implicit confermation of what the system has understood. Usually the system leaves free initiative to the user; only when the LP fails several times the system can take the initiative by guiding the user towards the accomplishment of the task. In such a case and when a focused answer from the user is predicted (like a confirmation) the DM can command the AFE to switch from CSR to IWR mode.

2.6 Text-to-Speech Synthesis Section

The generated answer is sent to the final text-to-speech [14], synthesis stage, housed in the PC-SYN cabinet, which performs two kinds of computation: linguistic processing and synthetic signal processing.

Linguistic processing converts the ASCII input text into a phonetic prosodic representation, that is a sequence of phonemes and their associated prosodic parameters, such as fundamental frequency, phoneme duration and intensity. Starting from this representation, the signal processing module accesses the diphone dictionary and generates the synthetic signal. The commands between the SUN4 and the PC-SYN are bidirectional: as an example the PC-SYN notifies the SUN4 when the synthesis is terminated, whilst the SUN4 issues to the PC-SYN high priority command for stopping the synthesis, having recognized the utterance FINE (i.e stop).

2.7 The system controller

A state machine system controller acknowledges the incoming call, changes the active vocabulary according to the dialogue evolution, stops the synthesizer on user request, controls the update of the remote data base and orderly terminates the session when the user disconnects.

3 System characterization

3.1 General considerations

The full characterization of a speech dialogue system is an open issue [13]; at the moment we characterised only the joint performance of the AFE and of the LP. We acknowledge that this is a rather partial test, which does not stress the sequential behaviour of the machine due to the dialogue module; however, we think that if the AFE+LP accuracy is not adequate, no dialogue recovery policy will be satisfactory.

We use 600 sentences for testing in the chosen application, read through the telephone by 10 different speakers: each of the speaker uttered a different set of 60 sentences. Suitable tools allow to record the samples of these utterances and to play back later as required.

Three alternatives are possible:

- the sentence is understood,
- the sentence understanding fails,
- the sentence is misunderstood, with two subcases: not recoverable and potentially recoverable.

We state that a sentence is understood by the LP if the extracted sentence is equal to the uttered one or possibly differs only for those short function words that our parser is able to skip during the analysis; in this case the sentence is also understood if these words are not essential: this happens almost always for the E-mail application.

The sentence goes failed in the case that the LP cannot extract any sentence within the available time and memory constraints, whilst the sentence goes misunderstood, if the LP extracts a wrong one. In this last case, if the LP looks for a set of solutions, there are two possibilities: the correct sentence is in the set, but it is not the chosen one (potentially recoverable case), or the correct sentence is not in this set (not recoverable case). A set of potentially recoverable sentences is eventually recognized after the feedback verification.

The percentage of recognised, failed, not recognized sentences depends on the LP time out: if less time is allowed, the number of failed sentences increases, whilst the number of recognized and not recognized sentences decreases; the opposite arises if more LP time is allowed. There is hence a trade-off between system accuracy and speed.

Failed and misunderstood sentences have not the same effect: in fact, in the failed sentences the system is aware of not having understood the question and can activate a recovery action through the dialogue manager, whilst if the sentence is misunderstood the system is not aware of its fault, at least at the LP level.

3.2 Specific measurements

Tab.1 summarizes the improvements we obtained in AFE+LF performance, mainly by improving the coupling between these subsystems. In the baseline system in fact the AFE forwarded to the LP only the word lattice: the

overall recognition rate was not satisfactory, though the two blocks were adequate when characterized separately.

A first improvement has been obtained by normalizing the lattice: in fact the Viterbi extracted lattice exhibits definite differences from frame to frame in the scores of the best words, differently from the Forward extracted lattices, which gave better results. We made the hypothesis that this difference in lattices could explain the different results and we normalized the Viterbi lattices by shifting all the scores in a frame in such a way that the best words are normalized to the same score value: the correctly understood sentences increased by a 10%. A second improvement has been the addition of a Word Penalty factor as a fixed cost for each new word hypothesized, to reduce the number of spurious short words appearing in the lattice.

A third minor improvement has been obtained with the addition of some word descriptions in the AFE, for composite words, and by adding the blow model: this result is reported in the column listed as "other updatings".

A fourth improvement has been obtained when the LP accounts for the best sequence of words besides the lattice: although the best sequence alone is not too reliable in our DDHMM Viterbi AFE, while it provides more correct information with Continuous Density HMMs [6], still it adds some useful information, which increases the number of understood utterances and decreases the misunderstood ones; moreover the LP computational time is further reduced.

The final improvement has been obtained by adding the feedback verification, which reduces recoverable misunderstood sentences allowing a correct understanding for more than half of them. This improvement has the further beneficial effect to validate short function words, but in any case increases computational times.

	base	normal latt.	penalty no ver.	other no ver.	+best sent. no ver.	+best sent. verify
Recogn.	48.8%	58.3%	62.3%	63.3%	67.2%	72.7%
Failed	18.8%	12.8%	5.2%	4.4%	4.2%	5.0%
Unrecog.	32.4%	28.9%	32.5%	32.3%	28.7%	22.3%
Recover.		15.8%	12.3%	12.5%	12.7%	4.7%
LP time		1.4 s.	0.66 s.	0.66 s.	0.44 s.	

Table 1: Summary of key improvements with the DDHMM Viterbi real-time AFE

4 Conclusions

First results obtained in our E-mail demonstrator are encouraging, although we know that further improvements are needed. In the future we aim at increasing the AFE+LP accuracy by upgrading the real-time demonstrator with improved features experimented in the off-line simulated recognizer; moreover, we will stress more the system robustness issues, required for dealing with the spontaneous speech of real users.

A system size reduction will be also considered, since it is now technically possible to implement all the system in a SUN4 with DSP accelerator boards.

References

[1] J. Peckham, *Speech Understanding and Dialogue over the Telephone: an Overview of the Esprit Sundial Project*, Proc. 4th DARPA Workshop on Speech and Natural Language, Pacific Grove, February 1991.

[2] A. Ciaramella, D. Clementino, R. Pacifici,*A PC Housed Speaker Independent Large Vocabulary Continuous Telephonic Speech Recognizer*, Proc. Eurospeech '91, Genova, September 1991.

[3] L. Fissore, P. Laface, G. Micca, R. Pieraccini, *Lexical Access to Very Large Vocabularies*, IEEE trans. ASSP, Vol. 37. N.8, August 1989, pp. 1197-1213.

[4] Kai Fu Lee, *Context dependent Phonetic Hidden Markov Models for Speaker-independent Continuous Speech Recognition*, IEEE trans. ASSP, Vol.38, n.4, April 1990, pp. 599-609.

[5] G. Pirani editor, *Advanced Algorithms and Architectures for Speech Understanding*, Research Reports ESPRIT-Project 26-SIP-Vol.1, Springer-Verlag 1990; chapter 3.2.

[6] L. Fissore, P. Laface, G. Micca, *Comparison of Discrete and Continuous HMMs in a CSR Task over the Telephone*, ICASSP '91, Toronto, Canada, 14-17 May, pp. 253-256.

[7] L. Fissore, P. Laface, G. Micca, R. Pieraccini, *Performance of a Speaker Independent Continuous Speech Recognizer*, Proc. NATO Workshop on Speech Recognition and Understanding, Cetraro (IT), 1-13 July 1990.

[8] E. Giachin, C. Rullent, *Linguistic Processing in a Speech Understanding System*, Proc. NATO Workshop on Speech Recognition and Understanding, Cetraro(Italy), 1-13 July 1990.

[9] M. Poesio, C. Rullent, *Modified Caseframe Parsing for Speech Understanding Systems*, Proc. IJCAI-87, Milano, August 1987.

[10] P. Baggia, E. Gerbino, E. Giachin, C. Rullent, *Efficient Representation of Linguistic Knowledge in Continuous Speech Understanding*, Proc. IJCAI-91, Sydney, August 1991.

[11] E. Giachin, C. Rullent, *Robust Parsing of Severely Corrupted Spoken Utterances*, Proc. COLING-88, Budapest, August 1988.

[12] E. Gerbino, P. Baggia, *Interpretation of Context-Dependent Utterances in Man-Machine Dialogue*, Proc. Eurospeech '91, Genova, Italy, September 1991.

[13] D.Pallett, *DARPA Resource Management and ATIS Benchmark Test Poster Session*, Proc. 4th DARPA Workshop on Speech and Natural Language, Pacific Grove, February 1991.

[14] L. Nebbia, *Text-to-Speech Synthesis System for Italian: an Overview*, Proc. of Verba90, Rome (Italy), Jan. 1990.